Л.В. Уланова

РАСПОЗНАВАНИЕ СИМВОЛОВ В ДРЕВНИХ МАНУСКРИПТАХ С ПОМОЩЬЮ ЕДИНСТВЕННОГО ЭКЗЕМПЛЯРА УЧЕБНОЙ ВЫБОРКИ

(University of California, Riverside, USA)

Интерес к прошлому был достаточно высок во все времена, и, как известно, одними из самых ценных источников древнего знания были письменные источники. До наших дней сохранилось немало древнейших документов, которые сейчас привлекают особо пристальное внимание, т.к. современные технологии позволяют разместить в сети огромное количество фотокопий древнейших манускриптов для всеобщего доступа [1 – 3]. В связи с этим на текущий момент немало научных работ посвящены исследованию и распознаванию древних рукописных и печатных документов.

Обычные техники сканирования и распознавания документов (Optical Character Recognition Systems – OCR) в данном случае работают достаточно плохо в связи с тем, что каждый документ имеет свой, порой уникальный, способ написания символов, а кроме того, если возраст документа насчитывает несколько столетий, материал манускрипта подвергается деградации. Это отчётливо видно на примере средневековой энциклопедии Liber Floridus [4, 5], фрагмент первой страницы которой представлен на рисунке 1.

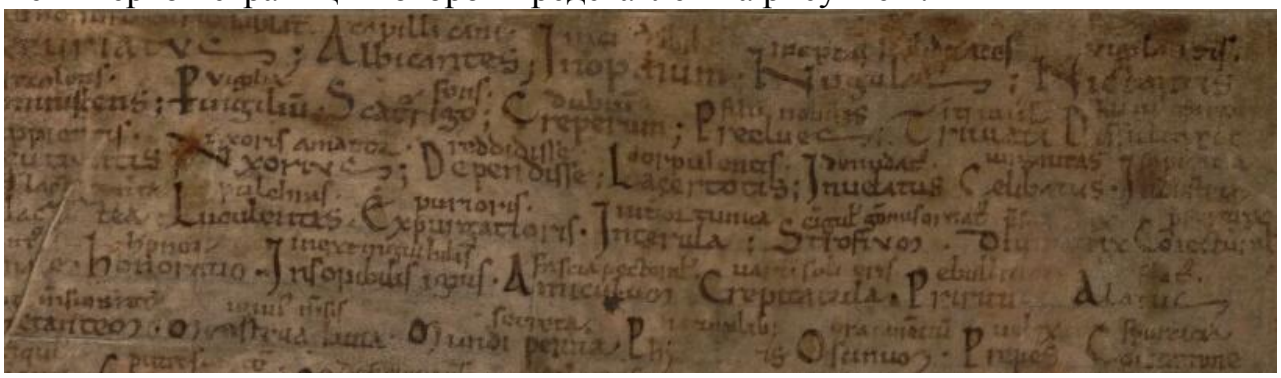

Рис. 1. Фрагмент первой страницы средневековой энциклопедии Liber Floridus

Поэтому очень часто для обработки и распознавания древних манускриптов применяется человечески труд, стоимость которого в разы превышает любую машинную обработку. Исследователи добились значительных результатов в компьютерных системах, работающих в диалоговом режиме, где программная система проводит предварительную классификацию, а затем представляет на суд пользователя наиболее сложные случаи. Например, такой подход применяется в [6, 7]. Авторы этих работ анализируют свойства конкретного документа и с помощью автоматизированных средств проводят его распознавание и анализ.

Но даже применяя работу вручную, не всегда получается добиться значительных результатов. Другой проблемой является недостаточный объём учебной выборки для каждого конкретного символа, т.к. для классификации неиз-

вестного символа нужно как можно больше примеров способов его написания. Для расширения выборки часто применяется техника, называемая Semi-Supervised learning (SSL). Суть этой техники заключается в том, что для классификации применяется не только выборка данных, предварительно подготовленных и классифицированных человеком, но и данные, которые были классифицированы программной системой в ходе распознавания, в нашем случае, символов текста древнего манускрипта. Таким образом получается расширить учебную выборку за счёт новых, только что полученных в ходе классификации, данных.

В целом, задача классификации мультимедийных данных сложнее, чем данных, которые можно описать набором свойств и их значений, представленным с помощью N-мерного вектора действительных чисел. Во втором случае речь идёт об анализе числовых данных, различие между которыми можно посчитать численно, применив некоторую функцию. В случае же мультимедийных, графических данных древних манускриптов, выразить свойства каждого символа помощью некоторого численного вектора крайне сложно. Одним из подходов является описание изображения с помощью контурной гистограммы [9, 10] с дальнейшим применением нейронных сетей для классификации.

Однако, данный метод был показан менее эффективным, чем анализ текстур [11]. Предложенный метод СК-1 (названный по фамилиям авторов: Campana-Keogh) показал высокую эффективность и точность в классификации и распознавании изображений в самых различных областях, в том числе, и анализе древних манускриптов [12]. Данный метод основан на теоретическом концепте колмогоровской сложности. Колмогоровская сложность строки $K(x)$ определена как длина кратчайшей программы, способной произвести x на универсальном компьютере (таком, как машина Тьюринга). Чтобы определить различие двух изображений, основываясь на принципе колмогоровской сложности, было рассмотрено несколько другое толкование этого понятия. Условная сложность $K(x|y)$ определена как длина кратчайшей программы, которая может вычислить x, если y является входным параметром. Другими словами, авторы описывают свою метрику как сравнение условных сложностей $K(x|y)$ и $K(y|x)$ к $K(xy)$, где последнее – длина программы, которая может вывести конкатенированную строку xy. Метрика может быть описана выражением (1):

$$d_k(x,y) = \frac{K(x|y) + K(y|x)}{K(xy)} \quad (1)$$

К сожалению, колмогоровскую сложность практически невозможно выразить численно, поэтому авторы [11] применили аппроксимацию и при помощи алгоритма сжатия данных, такого как использован в MPEG. Таким образом, для примера, можно представить, что СК-1 вычисляется как разница между двумя фреймами некоторой киноплёнки, где один фрейм – это первое изображение из учебной выборки, а второй – изображение, класс которого необходимо определить.

Метрика СК-1 появилась сравнительно недавно, поэтому на текущий момент ведутся работы по оптимизации применения её в различных областях. Ав-

тор данной статьи занимается изучением и анализом применения СК-1 в области исторических манускриптов для лучшего распознавания древних символов. В частности, изучением возможностей классификации символов, имея в учебной выборке малое число экземпляров.

Литература

1. Google Books Library Project,
2. http://www.google.com/googlebooks/library.html
3. Million Book Project (or the Universal Library), http://www.ulib.org/
4. The Centre d'Études Supérieures de la Renaissance, http://cesr.univ-tours.fr/
5. Derolez A., Lamberti S. Audomari canonici liber floridus: Codex autographus bibliothecae universitatis gandavensis, Ghent, 1968
6. Liber Floridus on-line http://www.liberfloridus.be/
7. Asi A., Rabaev I., Kedem K., El-Sana J. User-assisted alignment of arabic historical manuscripts // Proceedings of the 2011 Workshop on Historical Document Imaging and Processing (HIP`11), 2011. – pp. 22 – 28
8. Rabaev I., Biller O., El-Sana J., Kedem K., Dinstein I. Case study in Hebrew character searching // International Conference on Document Analysis and Recognition (ICDAR2011), 2011. – pp. 1080 – 1084
9. Chapelle O., Schölkopf B., Zien A. Semi-supervised learning. Cambridge, Mass.: MIT Press, 2006
10. Ha T.M., Bunke H. Handwritten numeral recognition by perturbation method // Proc. Fourth Int'l Workshop Frontiers of Handwriting Recognition, Taipei, Taiwan, Dec. 7-9, 1994. – pp. 97 – 106
11. Ha T.M., Bunke H. Design, implementation, and testing of perturbation method for handwritten numeral recognition // Technical Report IAM-96-014, Inst. of Computer Science and Applied Mat., University of Berne, Switzerland, 1996
12. Campana B., Keogh E. A compression based distance measure for texture // Proceedings of the 2010 SIAM International Conference on Data Mining (SDM'10), 2010. – pp. 850 – 861
13. Hu B., Rakthanmanon T., Campana B., Mueen A., Keogh E. Image mining of historical manuscripts to establish provenance // Twelfth SIAM International Conference on Data Mining (SDM 2012), 2012